REFORECASTS, AN IMPORTANT NEW DATA SET FOR IMPROVING WEATHER PREDICTIONS

Thomas M. Hamill, [1] Jeffrey S. Whitaker, [2]
and Steven L. Mullen [3]

[1] *University of Colorado and NOAA-CIRES Climate Diagnostics Center, Boulder, CO*

[2] *NOAA-CIRES Climate Diagnostics Center, Boulder, CO*

[3] *Institute of Atmospheric Physics, University of Arizona, Tucson, AZ*

7 March 2005

Submitted to *Bulletin of the American Meteorological Society*

Corresponding author address: Dr Thomas M Hamill, NOAA-CIRES Climate Diagnostics Center, Boulder, CO 80305-3328. E-mail: tom.hamill@noaa.gov
Phone 1 (303) 497-3060

ABSTRACT

A "reforecast" (retrospective forecast) database has been developed. This database is comprised of a 15-member ensemble of a T62 forecast model run out to two weeks lead. Forecasts have been run every day from 0000 UTC initial conditions from 1979 to present. The model is a 1998 version of the National Centers for Environmental Prediction's Global Forecast System (NCEP GFS) at T62 resolution. The 15 initial conditions consist of a reanalysis and seven pairs of bred modes.

This database facilitates a number of applications that were heretofore impossible. Systematic model errors can be diagnosed from the past forecasts and corrected, thereby dramatically increasing forecast skill. For example, calibrated precipitation forecasts over the United States based on the 1998 reforecast model are much more skillful than precipitation forecasts from the current, higher-resolution version of the NCEP GFS. Other applications are also demonstrated, such as the diagnosis of bias and an identification of the most predictable patterns of week 2 forecasts.

It is argued that the benefits of reforecasts are so large that they should become an integral part of the numerical weather prediction process. Methods for integrating reforecast approaches without seriously compromising the pace of model development are discussed.

Users wishing to explore their own applications of reforecasts can download the real-time and retrospective forecasts through a web interface.

1. Introduction

Reanalyses such as the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis (Kalnay et al. 1996) and the European Centre for Medium Range Weather Forecasts (ECMWF) 40-year reanalysis (ERA-40; Uppala et al. 2004) have become heavily used products for geophysical science research. These reanalyses run a practical, consistent data assimilation and short-range forecast system over a long period of time. While the observation type and quality may change somewhat, the forecast model and assimilation system are typically fixed. This facilitates the generation of a reanalysis data set that is fairly consistent in quality over time. The reanalysis data set has facilitated a wide range of research; for example, the Kalnay et al. article above has been cited more than 2700 times at the time of writing.

In this article we explore the value of a companion data set to reanalyses, which we shall call "reforecasts." These are retrospective weather forecasts generated from a fixed numerical model. Model developers could use them for diagnosing model bias, thereby facilitating the development of new, improved versions of the model. Others could use them as data for statistically correcting weather forecasts, thereby developing improved, user-specific products (e.g., Glahn and Lowry 1972, Carter et al. 1989). Others may use them for studies of atmospheric predictability. Unfortunately, extensive sets of reforecasts are not usually produced. These computationally expensive reforecasts are "squeezed out" by operational data assimilation and forecast models run at as fine a resolution as possible.

Would the additional forecast improvement and diagnostic capability provided by reforecasts make them worth the extra computational resource they require? To explore

this, we recently generated a prototype 25-year, 15-member ensemble reforecast data set using a 1998 version of the NCEP MRF model run at T62 resolution – admittedly a resolution far from state-of-the-art in operational numerical weather prediction centers in 2005. Despite the coarse resolution of this data set, we were able to make probabilistic week 2 forecasts that were more skillful than the operational NCEP forecasts based on higher-resolution models (Hamill et al. 2004). Vitart (2004) also demonstrated improved monthly forecasts using a smaller reforecast data set, and the "DEMETER" project in Europe has tested the approach in a multi-model environment for seasonal climate forecasts (Palmer et al. 2004, Hagedorn et al. 2005, Doblas-Reyes et al. 2005). Given these encouraging results, we have begun to explore other applications to reforecasts.

Our intent in this article is introduce the reader to the several applications of reforecast data that demonstrate the potential for improving weather predictions and increasing our understanding of atmospheric predictability. We also intend to stimulate a serious discussion about the value of reforecasts. Is the value added so large that operational weather forecast centers should make reforecasting a regular part of the operational numerical weather prediction process? We will demonstrate that there is a large, additional amount of forecast skill that can be realized through the use of the reforecasts. Because reforecasting using higher-resolution models is expensive, implementing this idea would require the purchase or reallocation of computer resources. Thus the implementation of reforecasting requires discussion at the top levels of the weather services.

We will provide a description of this database in section 2 and illustrate how users can download raw data. We then demonstrate a range of potential applications in section

3, illustrating how such databases may be used to improve a variety of forecast problems. Section 4 provides conclusions and a discussion of the wider issues of reforecasting.

2. **Description of the reforecast data set.**

A T62 resolution (roughly 200 km grid spacing) version of NCEP's Global Forecasting System (GFS) model (Kanamitsu 1989; Kanamitsu et al. 1991; Hong and Pan 1996, Wu et al. 1997, Caplan et al. 1997, and references therein) was used with physics that were operational in the 1998 version of the model. This model was run with 28 vertical sigma levels.

A 15-member ensemble was produced every day from 1979 to current, starting from 0000 UTC initial conditions. The ensemble initial conditions consisted of a control initialized with the NCEP-National Centers for Atmospheric Research (NCAR) reanalysis (Kalnay et al. 1996) and a set of 7 bred pairs of initial conditions (Toth and Kalnay 1993, 1997) re-centered each day on the reanalysis initial condition. The forecasts extend to 15 days lead time, with data archived every 12 h.

Because of the large size of this data set, we have chosen to archive only a limited set of model output. Winds, temperature, and geopotential height are available at the 1000, 850, 700, 500, 250, and 150 hPa levels. 10-m wind components, 2-m temperature, mean sea-level pressure, accumulated precipitation, convective heating, precipitable water, and 700 hPa relative humidity were also archived. Data can be downloaded using the online web form http://www.cdc.noaa.gov/reforecast/ (Fig. 1). Real-time data can also be ftp'ed from ftp://ftp.cdc.noaa.gov/Datasets.other/refcst/ensdata/yyyymmddhh, where yyyy is the year, mm is the month, dd is the day, and hh is the hour of the

initialization time. The real-time forecasts are typically available about 10 hours after initialization time.

3. **Some applications of the reforecast data set**.

As reanalyses have fostered many creative diagnostic studies, a long reforecast data set permits an examination of weather forecasts in ways that were not previously possible. Robust statistical forecast techniques can be developed, the characteristics of model biases more thoroughly understood, and predictability issues explored. We demonstrate here some interesting applications of reforecasts.

a. Forecasting with observed analogs.

Many forecast users desire reliable, skillful *high-resolution* ensemble predictions, perhaps for such applications as probabilistic quantitative precipitation forecasting or hydrologic applications (e.g., Clark and Hay 2004). The data set produced in this pilot reforecast project is comparatively low resolution, T62. However, it may be possible to downscale and correct systematic errors in ensemble forecasts through analog techniques, producing a high-resolution probabilistic forecast. Given a long time series of reforecasts and high-resolution analyses or observations, a two-step procedure is invoked. First, today's ensemble forecast is compared to reforecasts of the same lead. Second, the dates of the closest pattern matches are noted, and an ensemble is formed from the observed or analyzed conditions on those dates.

This two-step procedure is appealing, for it simulates the forecast process of many humans: we look at the current forecast, recall situations where the forecast depiction was

comparable (step 1), and try to recall the weather that actually occurred (step 2). Analog forecast techniques have a rich history (e.g., Toth 1989, van den Dool 1989, Livezey et al. 1994, Zorita and von Storch 1999, Sievers et al. 2000), but most utilize a simpler approach of directly finding observed analogs to the forecast. Consider a situation where the forecast model is consistently too warm. In this one-step analog technique, the ensemble of observed analogs would, by construction, retain the forecast's warm bias. The two-step procedure would first find similar forecasts, but if the observed data were cooler, the second step would compensate for the warm bias.

To demonstrate the potential of this two-step analog procedure, the technique was used to generate probabilistic forecasts of 24-h accumulated precipitation over the conterminous United States (US). Forecasts were verified during January-February-March 1979-2003 (JFM 79-03). Approximately 30-km North American Regional Reanalysis (NARR; Mesinger et al. 2005) data was used both for verification and as the data set from which historical observed weather analogs were selected.

The first step of the procedure was to find the closest *local* reforecast analogs to the current numerical forecast. That is, within a limited-size region, today's forecast was compared against past forecasts in that same region and at the same forecast lead time. Specifically, the ensemble-mean precipitation forecast pattern was computed at a subset of 16 coarse-resolution grid points (for example, the blue dots in Fig. 2). Reforecasts for this region were compared over all years, but only for a window of 91 days (+ / - 45 day window) around the date of the forecast. For example, a 15 February 2002 4-day ensemble-mean forecast over the northwest US was compared against the 4-day ensemble mean reforecasts from 1 January – 1 April 1979 - 2001 over the northwest US.

The root-mean square (RMS) difference between the current forecast and each reforecast was computed, averaged over the 16 grid points in Fig. 1. The 75 historical dates with the smallest RMS difference were chosen as the dates of the analogs.

The second step was the extraction of the observed weather on the dates of the closest 75 analogs. For this application, the NARR observed precipitation states were then extracted at the interior red dots in Fig. 2. A probabilistic forecast was then generated using the ensemble relative frequency; for example, if 50 of the 75 members at a grid point had greater than 10 mm accumulated rain, the probability of exceeding 10 mm was set to 67%.

The process was then repeated for other locations around the US. A full, high-resolution probabilistic forecast was generated by tiling together the local analog forecasts.

Figure 3 shows the data from one such case, a 3-day heavy precipitation event along the west coast in late December, 1996 (Ralph et al. 1998, Ogston and Hay 2000). Figure 3a shows the probabilistic forecasts generated from the raw T62 ensemble. Regions where the ensemble forecast members exceeded 100 mm of rainfall over the two days excluded Washington State, and the highest probabilities of greater than 100 mm were in a diffuse area in northern California. In comparison, when probabilities were computed from the 75 historical analogs, nonzero probabilities for exceeding 100 mm were extended north into Washington State. The high probability tended to be localized more along the coastal mountain ranges, the Cascades, and the Sierra Nevada range. The observed precipitation (Fig. 3c) shows that the heaviest precipitation had a similar spatial

pattern to the pattern of high probabilities, with heaviest precipitation along the mountain ranges.

To understand better how the analog technique works, consider the forecasts at the three dots in Fig. 3a. Figure 4a provides information for the grid point near Mt. Shasta, in northern California. The scatterplot shows the closest 75 ensemble-mean reforecast values of rainfall (abscissa) plotted against the associated 75 historical NARR analyzed rainfall values (ordinate). The histogram for the NARR analog ensemble is plotted along the right-hand side. The histogram along the top denotes the raw T62 ensemble forecast information, showing that the precipitation was exceptionally heavy in this area for all ensemble members. As indicated by the difference in the position of the forecast histogram and forecast analogs dots, the reforecast data was not able to find many close forecast analogs to this record event. However, the observed amounts associated with even these relatively poor analogs often indicated heavy precipitation, and the uncertainty of the forecast was increased, with NARR analog forecast members ranging from 0 to 324 mm.

In Fig. 4b, the raw forecast ensemble at Medford, Oregon also indicated a heavy precipitation event, again indicated by position of the histogram at the figure top. As with Mt. Shasta, no close analogs could be found, so the reforecast analogs were uniformly lighter in amount. However, this location was in a climatological rain shadow of the Coast Range in Oregon and California. The smoothed terrain in the reforecast model was unable to resolve this level of terrain detail, so heavier precipitation than observed was commonly forecast in Medford. Consequently, the two-step analog procedure corrected the overforecast bias.

In Fig. 4c, moderate precipitation amounts were forecast in the original ensemble in the Olympic Range of Washington State, and many similar reforecast analogs were found in the data set. The associated NARR observed analogs tended to be heavier in amount, with a larger spread than the original ensemble, thus correcting for an underforecasting bias and what was probably insufficient spread in the original ensemble.

It appears that the analog technique may ameliorate problems with bias and inadequate spread in the original ensemble. Still, does the skill of this technique exceed that from existing operational ensemble forecasts? To determine this, we have extracted the operational ensemble forecasts from NCEP for JFM 02-03; starting in January 2002, NCEP GFS ensemble forecasts were computed at T126 resolution to 84 h lead, providing it with a resolution advantage over the reforecast model. Figures 5a-b show the Brier Skill Score (Wilks 1995) of ensemble forecasts at the 2.5 mm and 25 mm (per 24 h) thresholds, respectively. The analog reforecast technique is much more skillful than the NCEP forecast, especially at the 25 mm threshold.

The increased skill of the analog forecasts relative to operational NCEP forecasts results suggest that there is some benefit from the use of analogs, the large 23-year training data set, or both. Figure 4 demonstrated how the use of high-resolution observed analogs permitted the extraction of small-scale detail that was not in the original forecast. But are two-plus decades of reforecasts necessary? Figure 6 indicates that forecast skill is degraded considerably when shorter training data sets are used, especially at high precipitation thresholds. In these cases, when a large amount of rain is forecast, it is important to have other similar high-rain forecast events in the data set, otherwise very

few close analogs can be found. For 2.5 mm forecast amounts, where there are many similar analogs in the reforecast database, and little skill is gained between 12 and 24 years of training data, indicating that the sample size of 12 years is probably adequate. However, for 25 mm, there still is a notable increase in skill between 12 and 24 years, indicating the importance of a long training data set.

Figure 7 shows the spatial pattern of the analogs' Brier Skill Score at 2.5 mm and 4 days lead. Skill varied greatly with location. The method produces highly skillful forecasts along the west coast, presumably because the methodology provides a way of downscaling the weather appropriate to the complex terrain. Forecast skill was generally high over the eastern US and lower over the northern US and Rockies. The mechanisms responsible for these large spatial variations in skill are not well understood (although the reforecast dataset itself provides a unique data for studying predictability questions such as this, see e.g. section 3c). Since the NARR analyzed precipitation is based primarily on gage and radar data, each of which are far less accurate in regions of snowfall than rain, these regional variations in skill may be partly a consequence of the observational data itself. In any case, the forecasts are generally quite reliable (Fig. 8), though less so at 1-day forecast leads (not shown).

Analog techniques must be used with caution. They will never predict record-setting events, events that lie outside the span of the past data. Also, as Lorenz (1969) noted, it is impossible to find *global* analogs for the current weather during a span of time as long as the recorded history of the atmosphere. Hence, analogs must be found and applied only in geographically limited regions, so that the difference between the current forecast and a past forecast analog is a small fraction of the climatological forecast

variance. Still, analog techniques should have very notable advantages. One notable benefit is that it should be conceptually easy to develop probabilistic weather forecasts tailored to a wide range of user concerns. To consider one example, suppose a user requires probabilistic ocean wave forecasts near an offshore oil well. If a large database of past observations of wave height at the oil well is available, the basic technique can be tailored to this application: find an ensemble of past forecast days where the meteorological wind conditions producing waves were similar to today's forecast, produce a probabilistic forecast from the associated observed wave heights on the days of the analogs. For another recent work on forecast analogs using this reforecast data set for hydrologic applications, see Gangopadhyay et al. (2004).

The analog technique demonstrated here is simply a proof of concept; perhaps analog forecasts could be improved by weighting closer analogs more than farther analogs, or perhaps the technique could be improved by finding multi-parameter analogs, such as matching winds, temperature, and precipitation forecasts rather than precipitation alone. Perhaps other statistical techniques will prove superior to this analog approach. Though there may be room for exploration and improvement, this simple algorithm in conjunction with low-resolution reforecasts was able to produce probabilistic forecasts greatly exceeding the skill of the higher-resolution NCEP global ensemble forecast system.

b. *Diagnosing model bias from reforecasts.*

Reforecasts are a useful tool for diagnosing model bias, the mean forecast minus the mean verification. For example, Figure 9a shows the bias of 850 hPa temperature

forecasts at a location near Kansas City. These biases were calculated by subtracting the NCEP-NCAR analyses from the ensemble-mean forecasts using 1979-2001 data and a 31-day window centered on the day of interest. Note the complex structure of the biases, from a large cold bias at shorter lead times in the winter and late-summer warm biases, especially at longer leads. Clearly, it would be mistaken to apply, say, a bias correction from early fall to late fall forecasts. And though not shown here, different locations have very different bias characteristics; for example, near San Francisco, there is a strong cold bias at longer leads during mid-summer.

Can the bias corrections be properly estimated from a much shorter data set of reforecasts? Figure 9b suggests that they cannot. This panel shows the standard deviation of the yearly bias estimates. To generate this figure, the bias was estimated for each year, day, and forecast lead using just a 31-day window centered on the day of interest. From the 23 bias estimates from 1979 to 2001, the standard deviation was calculated and plotted. Note that the standard deviation grows with increasing forecast lead and is generally larger in the winter than in the summer. This is due to the larger variability of the forecasts during the wintertime at longer leads. At most forecast leads and times of the year the spread in the yearly bias estimate is larger than the magnitude of the bias in Fig. 9a. This indicates that it is generally inappropriate to estimate model bias from a small sample of forecasts, and in fact the supposed "bias correction" will often actually introduce more error than it will correct. This points out the value of training statistical models with a large sample of forecasts; for another example for week 2 forecasts, see Hamill et al. (2004)

c. *Studying predictability using reforecasts*.

Forecasts of individual weather systems during the first week are generally referred to as "weather" forecasts, whose skill is primarily driven by sensitivity to initial conditions. Long-lead predictions, such as those associated with El-Nino/Southern Oscillation (ENSO) are generally referred to as "climate" forecasts, whose skill is primarily driven to sensitivity to boundary conditions (e.g. sea-surface temperatures). In between these two extremes lies the boundary between weather and climate forecasts, in which individual weather systems may not be predictable, but larger-scale flow patterns which influence those weather systems may retain some sensitivity to initial conditions, and may also be influenced by persistent boundary forcing. The phenomena that yield skill in the second week of an ensemble forecast are generally large scale and low frequency, and hence there may be only a few independent samples of these events each season. In addition, the predictable signal may be small compared to the uncertainty in a single forecast, so ensembles may be needed to extract that signal. Quantifying the nature of the predictable signal in week two therefore requires a large sample of ensemble forecasts, spanning many years. The reforecast dataset is the first such dataset to satisfy these requirements. Very basic questions, like "how much skill is there in week two?" and "where does that skill come from?" remain largely unanswered. In this section we show a few simple diagnostics using the reforecast dataset which provide some insight into these questions They are primarily intended to illustrate the utility of the reforecast dataset in investigations of atmospheric predictability.

Figure 10 shows a map of the temporal correlation between the time series of ensemble mean forecast and observed 500 hPa height for all day 10 forecasts in the

reforecast dataset initialized during December-February 1979-2003. Values locally exceeded 0.6 in the central Pacific, while the hemisphere average was 0.47. While these values may seem low, they do indicate that skillful probabilistic forecasts is possible at day 10. Hamill et al. (2004, Fig. 6d) showed that a correlation of 0.5 can be translated into a Ranked Probability Skill Score of 0.15 for terciles of the climatological probability distribution; small, but certainly useful.

What are the skillfully predicted patterns in these day-10 forecasts? To answer this question, we have performed a statistical analysis of the 25-year dataset of wintertime day-10 500 hPa height forecasts in order to identify the most predictable patterns. The technique used was Canonical Correlation Analysis (CCA; Bretherton et al., 1992), which seeks to isolate the linear combination of data from a predictor field and the linear combination of data from a predictand field that have the maximum linear correlation. The analysis was performed in the space of the truncated principal components, or PCs (Barnett and Preisendorfer, 1987). Here the predictor field consisted of the leading 20 PCs of ensemble mean week two forecast 500 hPa height, while the predictand field consisted of the leading 20 PCs of the corresponding weekly mean verifying analyses. The analysis was similar in spirit to that performed by Renwick and Wallace (1985), using 13 years of day-10 forecasts from the US National Weather Service. Although their 'most predictable pattern' is very similar to the one shown here, interpretation of their results was hampered by the fact that the forecasts were from a single deterministic run (not an ensemble) and there were significant changes in the forecast model over the span of their forecast archive.

Figure 11 shows the three most predictable patterns identified by the CCA analysis for day-10 forecasts, while Figure 12 shows the correlation between the time series of these patterns as a function of forecast lead. The patterns were computed for day-10 forecasts, but we have projected the forecasts for all other forecast leads on to these same patterns to see how the forecast skill evolves during the forecast period. The most predictable patterns are similar to well-known recurring persistent circulation anomalies, often called 'teleconnection patterns' (Barnston and Livezey, 1987). The first pattern is similar to the Tropical/Northern Hemisphere pattern, so named by Mo and Livezey (1986), which are often observed to appear in Northern Hemisphere wintertime seasonal means during El Nino/Southern Oscillation (ENSO) events. Indeed, the model forecast tropical precipitation during the first week regressed on the time series of this pattern (not shown) shows a significant relationship between precipitation in the central equatorial Pacific and the amplitude of the most predictable pattern at day 10. However, values of this correlation were less than 0.3, indicating that only a modest fraction of the variance of this pattern in day 10 forecasts is directly related to variations in tropical convection. The second most predictable pattern was similar to the Pacific/North American pattern (Wallace and Gutzler, 1981), while the third resembled the classical North Atlantic Oscillation (ibid). Regression analysis shows that neither of these patterns had a strong relationship with the model forecast precipitation in equatorial regions during week one, implying that dynamics internal to the extratropics was primarily responsible for their predictability. The ability of the ensemble system to forecast these patterns is remarkable, with correlation skill exceeding 0.7 at day 10 for all three patterns, and skill exceeding 0.6 at day 15 for the most predictable pattern.

These results show that the reforecast dataset provides a new opportunity to address basic predictability questions which were previously out of reach due to the sample size limitations and questions concerning the impact of model changes in existing operational forecast datasets.

4. **How can reforecasts be integrated into operational numerical weather prediction?**

This paper has demonstrated some of the benefits of "reforecasts," a companion database to reanalyses. As with reanalyses, a fixed model is used, and forecasts for retrospective cases are computed with the fixed model. This paper demonstrated that the use of such a data set can improve weather forecasts dramatically, can diagnose model biases, and provide enough forecast samples to answer some interesting questions about predictability in the forecast model.

Weather prediction facilities like the Environmental Modeling Center at NCEP already totally utilize the available computational resources, and planned model upgrades will utilize most of the newly available resources in the near future. How then can reforecasts be integrated into operational numerical weather prediction?

One possible interim solution is that the reforecasts could be run with a less than state-of-the-art version of the forecast model. Our results suggest that substantial forecast improvements were possible even with the T62 model output. Consider then the scenario where the "flagship" product at the weather production facility is a 50-level, 50-member global ensemble of a T300 model. The operational production of a 25-level, 25-member, T150 forecast could be produced at 1/64 the computational expense of the flagship

product. Hence, if 32 years of reforecasts were computed, this would require computer resources equivalent to producing the operational forecasts for half a year. If the companion reforecasts were computed offline on another computer, then the reforecast computations would barely affect operations. We suggest that this is an appropriate, conservative model to follow in the near future. A timely forecast would be generated from a fixed, reduced-resolution version of the model, one where a companion reforecast data set had been generated. Forecast products would be generated through statistical techniques such as those demonstrated here, and the reforecast-based products would be compared to products based on the flagship product. If they were deemed to improve weather forecast capabilities, then every few years the reforecast would be updated, utilizing a newer, improved version of the forecast model at higher resolution, maintaining the same relative usage of operational CPU resources relative to the updated flagship product.

Computing reforecasts is a task that is easily parallelized, so it can take advantage of massively parallel clusters of inexpensive computers. Individual ensemble member forecasts can be computed on different processors in the cluster, and forecasts for many different initial times can also be parallelized. The cluster of personal computers and storage array used to in this experiment cost approximately $90,000 (WHH04), a tiny fraction of the cost of the NCEP's supercomputers.

Acknowledgments

T. Hamill's contribution was supported under National Science Foundation grants ATM-0120154 and ATM-0205612. We thank Xue Wei and Andres Roubicek for their help in processing the voluminous data used in this study.

References

Barnston, A. G., and R. E. Livezey. 1987: Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Mon. Wea. Rev.,* **115**, 1083–1126.

Barnett, T. P., and R. Preisendorfer, 1987: Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Mon. Wea. Rev*., **115**, 1825–1850.

Bretherton C. S., C. Smith and J. M. Wallace. 1992: An intercomparison of methods for finding coupled patterns in climate data. *J. Clim*., **5**, 541–560.

Caplan, P., J. Derber, W. Gemmill, S.-Y. Hong, H.-L. Pan, and D. Parrish, 1997: Changes to the 1995 NCEP operational medium-range forecast model analysis-forecast system. *Wea. Forecasting*, **12**, 581-594.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting*, **12**, 581-594.

Clark, M. P., and L. E. Hay, 2004: Use of medium-range numerical weather prediction model output to produce forecasts of streamflow. *J. Hydrometeor*., **5**, 15-32.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting: Part II: calibration and combination. *Tellus*, in press.

Gangopadhyay, S., M.P. Clark, B. Rajagopalan, K. Werner, and D. Brandon, 2004: Effects of spatial and temporal aggregation on the accuracy of statistically

downscaled precipitation estimates in the Upper Colorado River basin. *Water Resources Research*, submitted.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203-1211.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: basic concept. *Tellus*, in press.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble re-forecasting: improving medium range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434-1447.

Hong, S.-Y., and H.-L. Pan, 1996: Nonlocal boundary layer vertical diffusion in a medium-range forecast model. *Mon. Wea. Rev.,* **124**, 2322-2339.

Kalnay, E., and co-authors, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, **77**, 437-472.

Kanamitsu, M., 1989: Description of the NMC global data assimilation and forecast system. *Wea. Forecasting*, **4**, 334-342.

---------- , and Coauthors, 1991: Recent changes implemented into the global forecast system at NMC. *Wea. Forecasting*, **6**, 425-435.

Livezey, R. E., A. G. Barnston, G. V Gruza, and E. Y. Ran'kova, 1994: Comparative skill of two analog seasonal temperature prediction systems: objective selection of predictors. *J Climate*, **7**, 608-615.

Lorenz, E.N., 1969: Three approaches to atmospheric predictability. *Bull. Amer. Meteor. Soc.*, **50**, 345-349.

Mo, K.C and R. E. Livezey. 1986: Tropical-extratropical geopotential height teleconnections during the northern hemisphere winter. *Mon. Wea. Rev*., **114**, 2488–2515.

Mesinger, F., G. DiMego, E. Kalnay, P. Shafran, W. Ebisuzaki, D. Jovic, J. Woollen, K. Mitchell, E. Rogers, M. Ek, Y. Fan, R. Grumbine, W. Higgins, H. Li, Y. Lin, G. Manikin, D. Parrish, and W. Shi, 2005: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc*., submitted.

Ogston, A., D. Cacchione, R. Sternberg, and G. Kineje, 2000: Observations of storm and river flood-driven sediment transport on the northern California continental shelf. *Continental Shelf Research*, **20**, 2141-2162.

Palmer, T. N., and others, 2004: Development of a European multimodel ensemble system for seasonal to interannual prediction (DEMETER). *Bull. Amer. Meteor. Soc.*, 85, 853-872.

Ralph, F.M., P.J. Neiman, P.O.G. Persson, and J.-W. Bao, 1998: Observations of California's New Year's Day Storm of 1997. *Preprints, 2nd AMS Conf. on Atmospheric and Oceanic Prediction and Processes*, 11-16 January, 1998, Phoenix, AZ, 219-224.

Renwick, J. A. and J. M. Wallace. 1995: Predictable anomaly patterns and the forecast skill of northern hemisphere wintertime 500-mb height fields. *Mon. Wea. Rev*., **123**, 2114–2131.

Sievers, O., K. Fraedrich, and C. Raible, 2000: Self-adapting analog ensemble predictions of tropical cyclone tracks. *Wea. Forecasting*, **18**, 3-11.

Toth, Z., 1989: Long-range weather forecasting using an analog approach. *J. Climate*, **2**, 594-607.

----------, and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.* **74**, 2317-2330.

----------, and ---------- , 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297-3319.

Uppala, S. M., and coauthors, 2004: The ERA-40 reanalysis. *Quart. J. Roy. Meteor. Soc.*, submitted.

van den Dool, H. M., 1989: A new look at weather forecasting through analogues, *Mon. Wea. Rev.*, **117**, 2230-2247

Vitart, F., 2004: Monthly forecasting at ECMWF. *Mon. Wea. Rev.*, **132**, 2761-2779.

Wallace, J. M. and D. S. Gutzler. 1981: Teleconnections in the geopotential height field during the northern hemisphere winter. *Mon. Wea. Rev.*, **109**, 784–812.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Cambridge Press, 467 pp.

Wu., W., M. Iredell, S. Saha, and P. Caplan, 1997: Changes to the 1997 NCEP Operational MRF Model Analysis/Forecast System. NCEP Technical Procedures Bulletin 443, 22 pp. [Available online at http://www.nws.noaa.gov/om/tpbpr.shtml, and from Office of Services, NOAA, National Weather Service, 1325 East-West Highway, W/0551, Silver Spring, MD, 20910.

Zorita, E., and H. von Storch, 1999: The analog method as a simple statistical downscaling technique: comparison with more complicated methods. *J. Climate*, **12**, 2474-2489.

List of Figures.

Figure 1: Screenshot of the reforecast dataset download web page.

Figure 2: Map of reforecast grid points used in determination of closest analog forecasts. The smaller dots denote where NARR data is available (a 32 km Lambert Conformal grid). Large blue dots denote where the T62 forecasts are available (interpolated from a global 2.5 degree grid to every eighth NARR grid point). The analyzed fields associated with the closest pattern matches at the blue dots are extracted at the red dots. The national forecast is then comprised of a tiling of similar regions from around the country.

Figure 3: (a) CDC raw ensemble-based probability of greater than 100 mm precipitation accumulated during days 4-6 for a forecast initialized 0000 UTC 26 December 1996 (from 0000 UTC 29 December 1996 to 0000 UTC 1 January 1997). (b) As in (a), but where probabilities have been estimated from relative frequency of historical NARR analogs. (c) Observed precipitation from NARR (mm). Dots indicate locations used in Fig. 4.

Figure 4: Ensemble forecast, reforecast analog, and observed analog data for three dots in Fig. 3a. Histograms along tops of plots indicate the raw T62 ensemble forecast total amounts. Histograms along right of plots indicate the frequency of NARR analog forecast amounts. Scatterplots indicate the joint value of ensemble-mean analog forecasts taken from the reforecast data set (abscissa) and the value of the associated NARR historical analog (ordinate). (a) Scatterplot from Mt. Shasta (northern California), (b)

scatterplot from Medford (southern Oregon), (c) scatterplot from Olympic Mountains, Washington.

Figure 5: Brier Skill Score of analog and NCEP ensemble forecasts measured relative to climatology. (a) 2.5 mm skill, (b) 25 mm skill.

Figure 6: Brier Skill Scores of the analog reforecast technique for various lengths of the training data set.

Figure 7: Map of Brier Skill Scores of 24-h accumulated precipitation forecasts between 3 and 4 days lead at 2.5 mm threshold for JFM 1979-2002.

Figure 8: Reliability of 2.5 mm probabilistic precipitation forecasts in Fig. 7.

Figure 9: (a) 850 hPa temperature bias at -95.0 W, 40.0 N, as a function of time of year and forecast lead. (b) Standard deviation of the yearly bias estimates.

Figure 10: Correlation between time series of ensemble mean day-10 forecasts and corresponding verifying analyses (from the NCEP/NCAR reanalysis) at every grid point in the Northern Hemisphere for December to February 1979-2003.

Figure 11 : Ensemble mean day-10 forecast 500 hPa height regressed on to the timeseries of the three most predictable forecast patterns. Contour interval 15 m. The

correlation of between the time series of the predictor pattern and the corresponding predictand pattern (r) is given for each pattern.

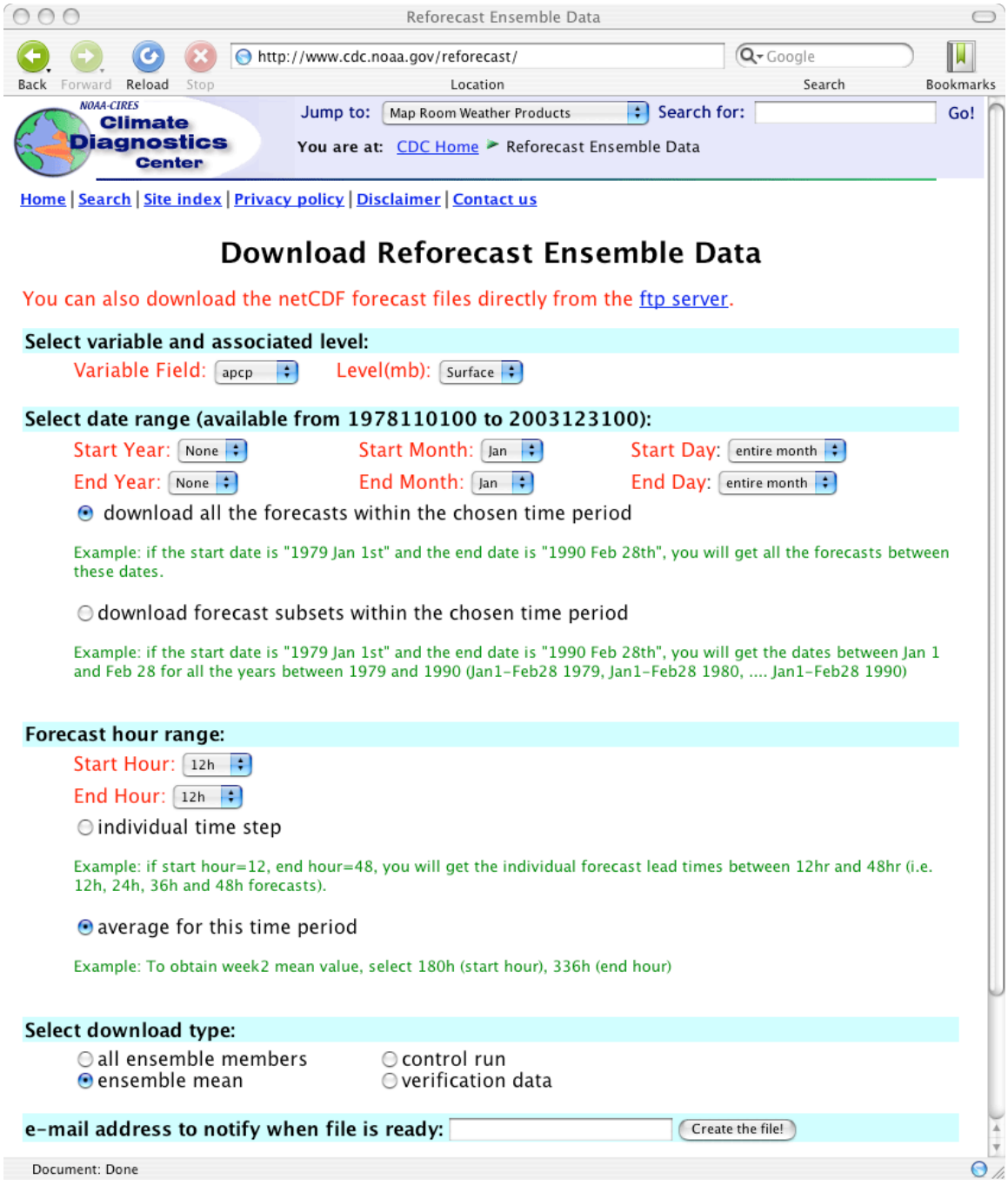

Figure 1: Screenshot of the reforecast dataset download web page.

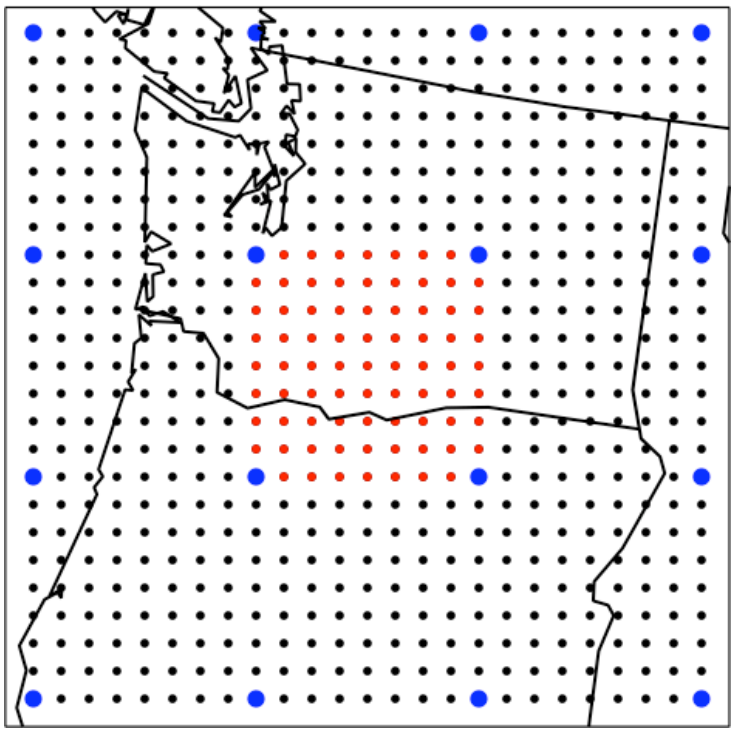

Figure 2: Map of reforecast grid points used in determination of closest analog forecasts. The smaller dots denote where NARR data is available (a 32 km Lambert Conformal grid). Large blue dots denote where the T62 forecasts are available (interpolated from a global 2.5 degree grid to every eighth NARR grid point). The analyzed fields associated with the closest pattern matches at the blue dots are extracted at the red dots. The national forecast is then comprised of a tiling of similar regions from around the country.

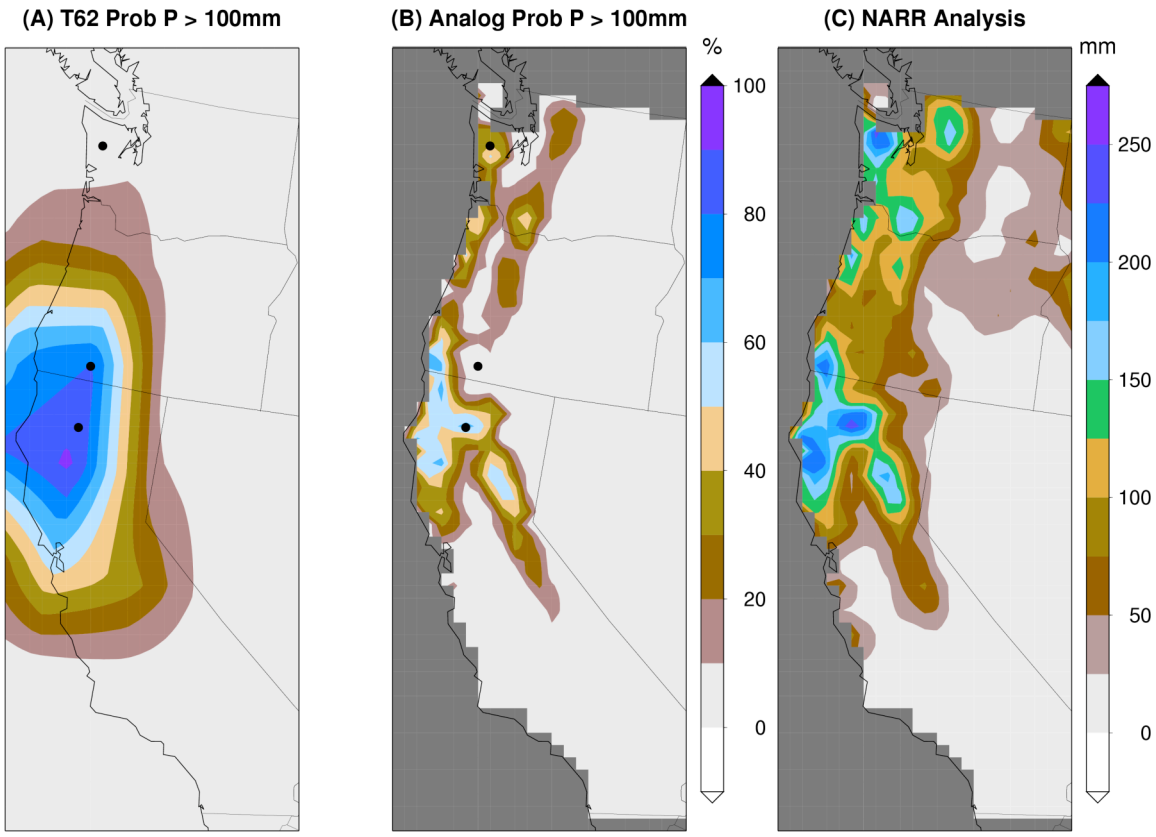

Figure 3: (a) CDC raw ensemble-based probability of greater than 100 mm precipitation accumulated during days 4-6 for a forecast initialized 0000 UTC 26 December 1996 (from 0000 UTC 29 December 1996 to 0000 UTC 1 January 1997). (b) As in (a), but where probabilities have been estimated from relative frequency of historical NARR analogs. (c) Observed precipitation from NARR (mm). Dots indicate locations used in Fig. 4.

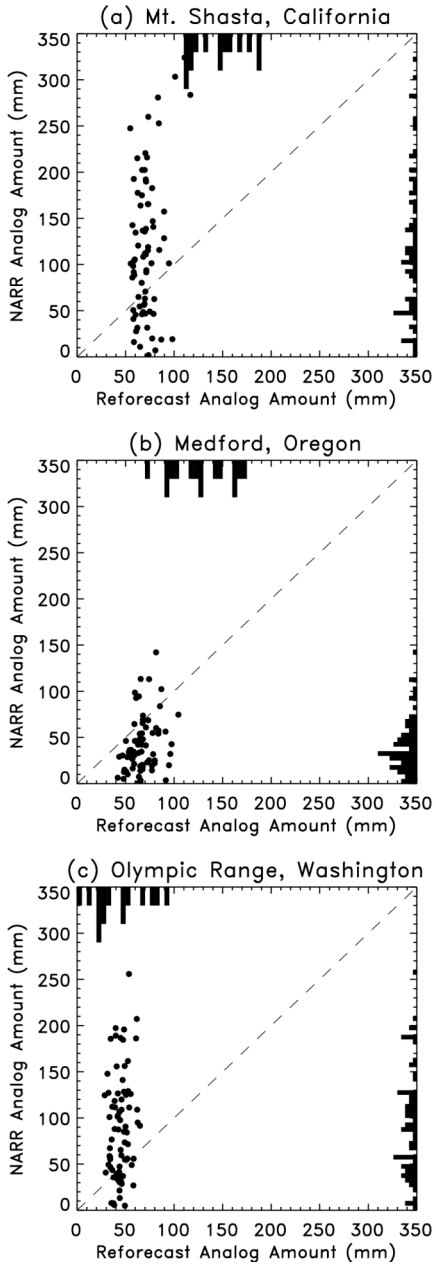

Figure 4: Ensemble forecast, reforecast analog, and observed analog data for three dots in Fig. 3a. Histograms along tops of plots indicate the raw T62 ensemble forecast total amounts. Histograms along right of plots indicate the frequency of NARR analog forecast amounts. Scatterplots indicate the joint value of ensemble-mean analog forecasts taken from the reforecast data set (abscissa) and the value of the associated NARR historical analog (ordinate). (a) Scatterplot from Mt. Shasta (northern California), (b) scatterplot from Medford (southern Oregon), (c) scatterplot from Olympic Mountains, Washington.

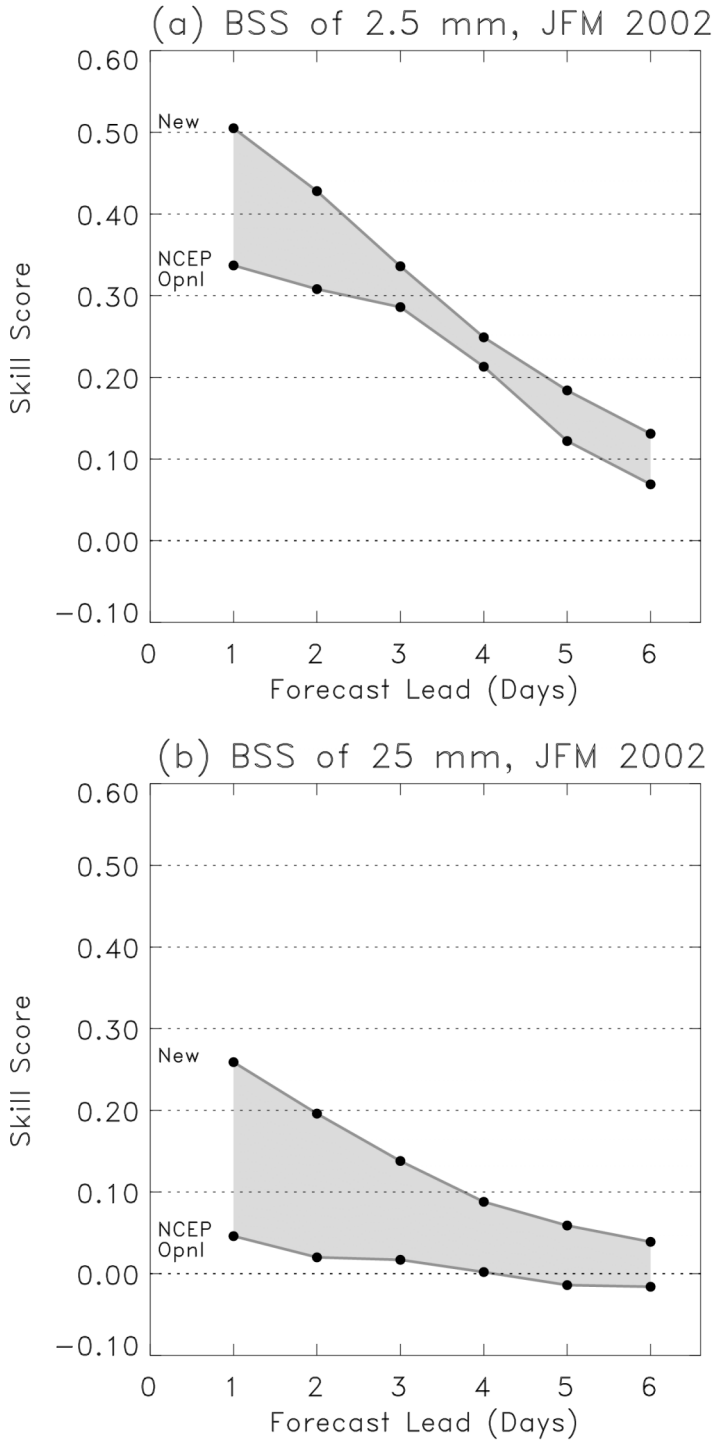

Figure 5: Brier Skill Score of analog and NCEP ensemble forecasts measured relative to climatology. (a) 2.5 mm skill, (b) 25 mm skill.

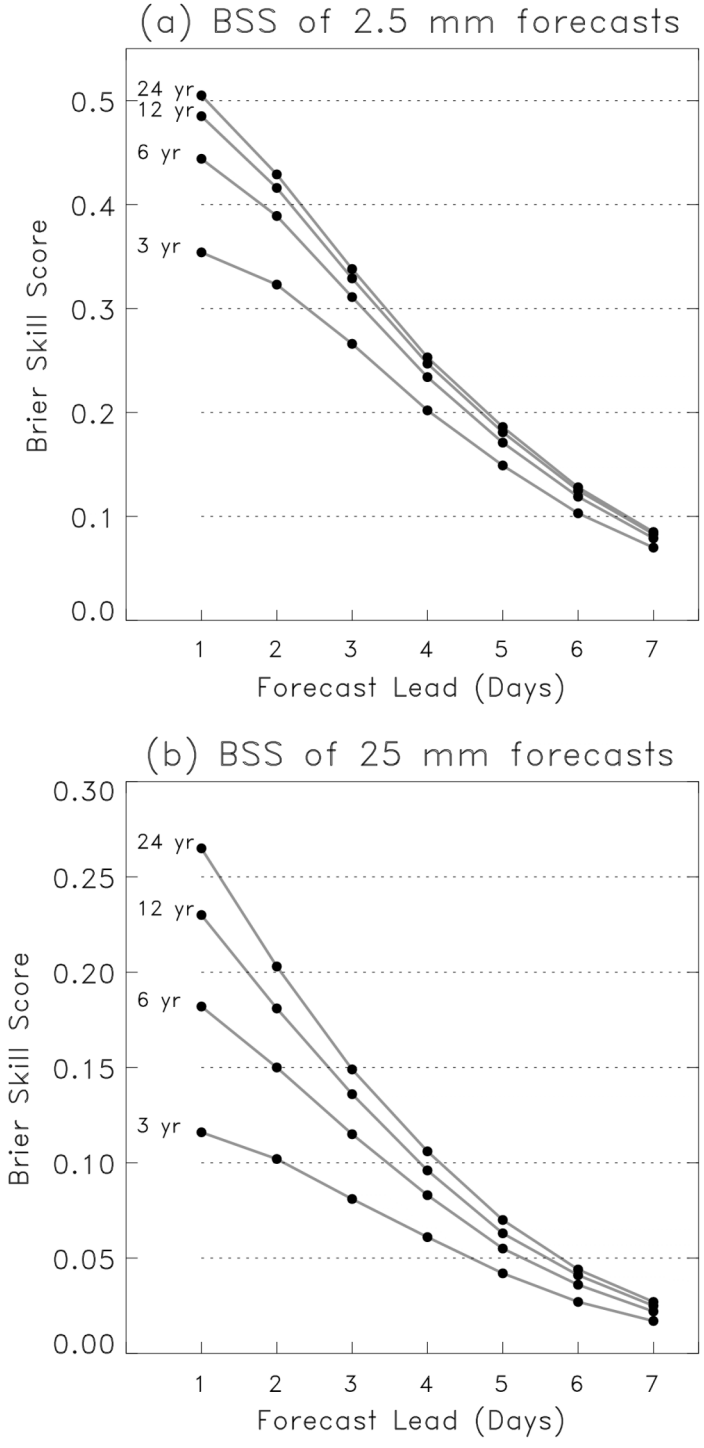

Figure 6: Brier Skill Scores of the analog reforecast technique for various lengths of the training data set.

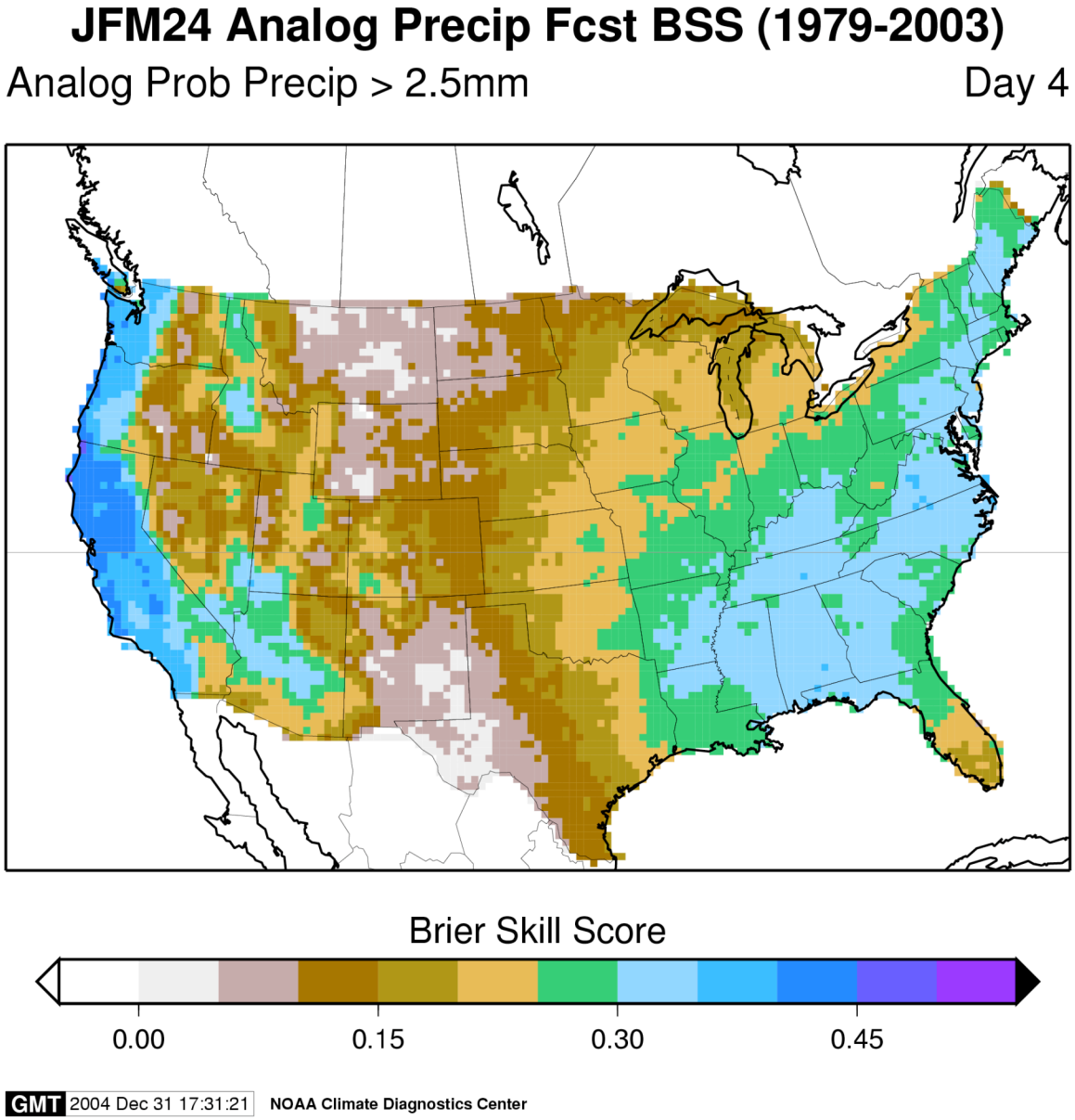

Figure 7: Map of Brier Skill Scores of 24-h accumulated precipitation forecasts between 3 and 4 days lead at 2.5 mm threshold for JFM 1979-2002.

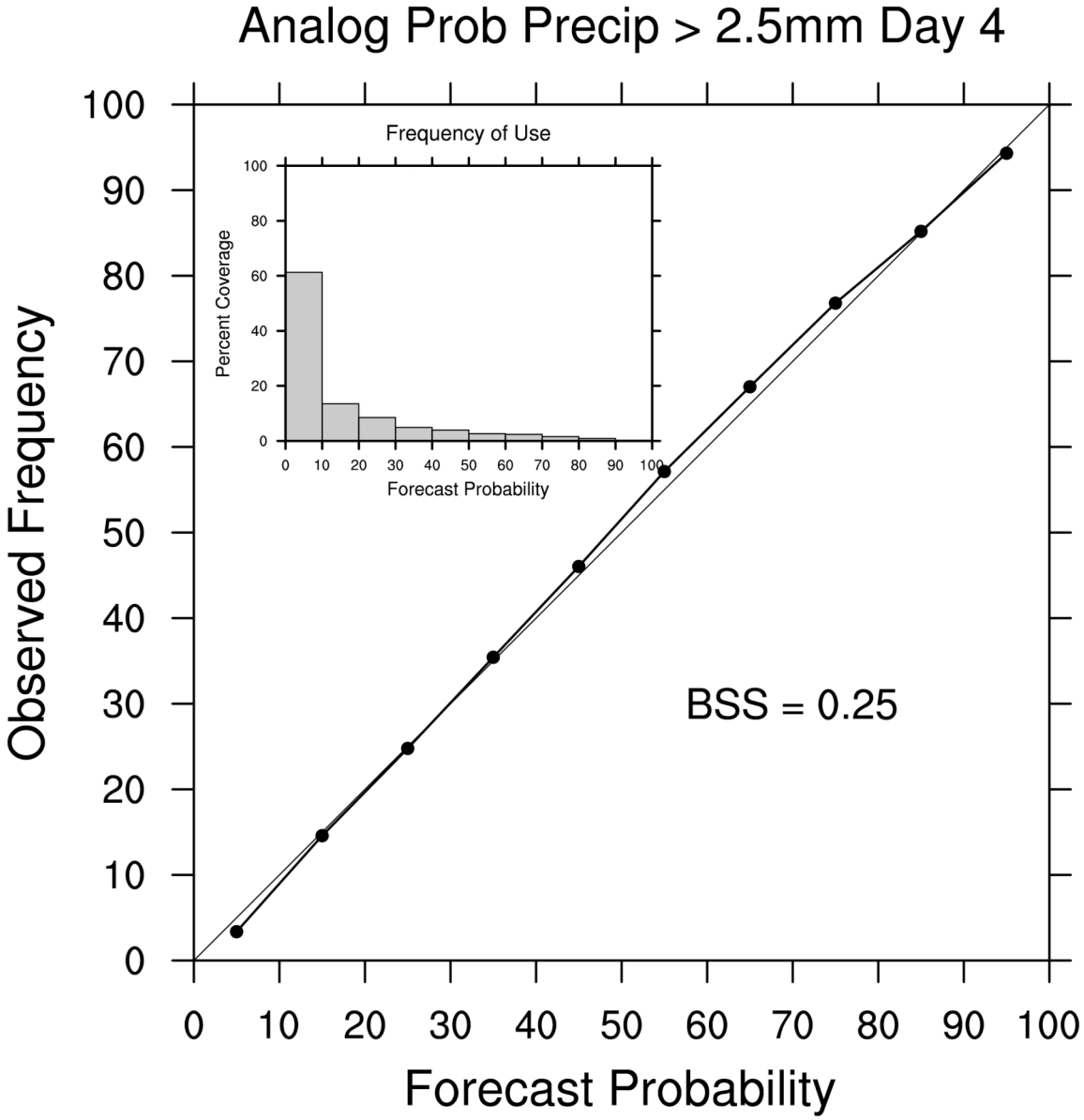

Figure 8: Reliability of 2.5 mm probabilistic precipitation forecasts in Fig. 7.

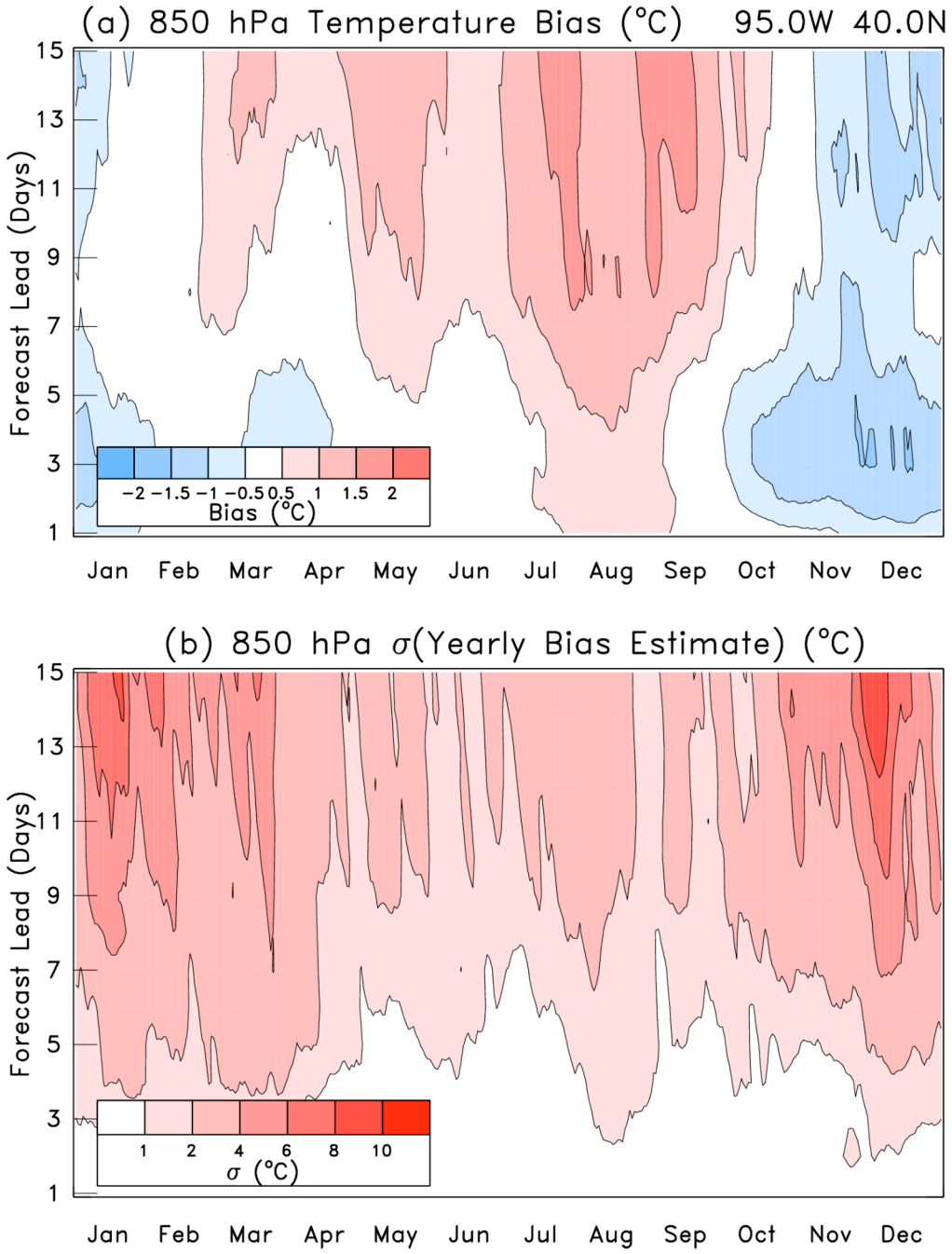

Figure 9: (a) 850 hPa temperature bias at -95.0 W, 40.0 N, as a function of time of year and forecast lead. (b) Standard deviation of the yearly bias estimates.

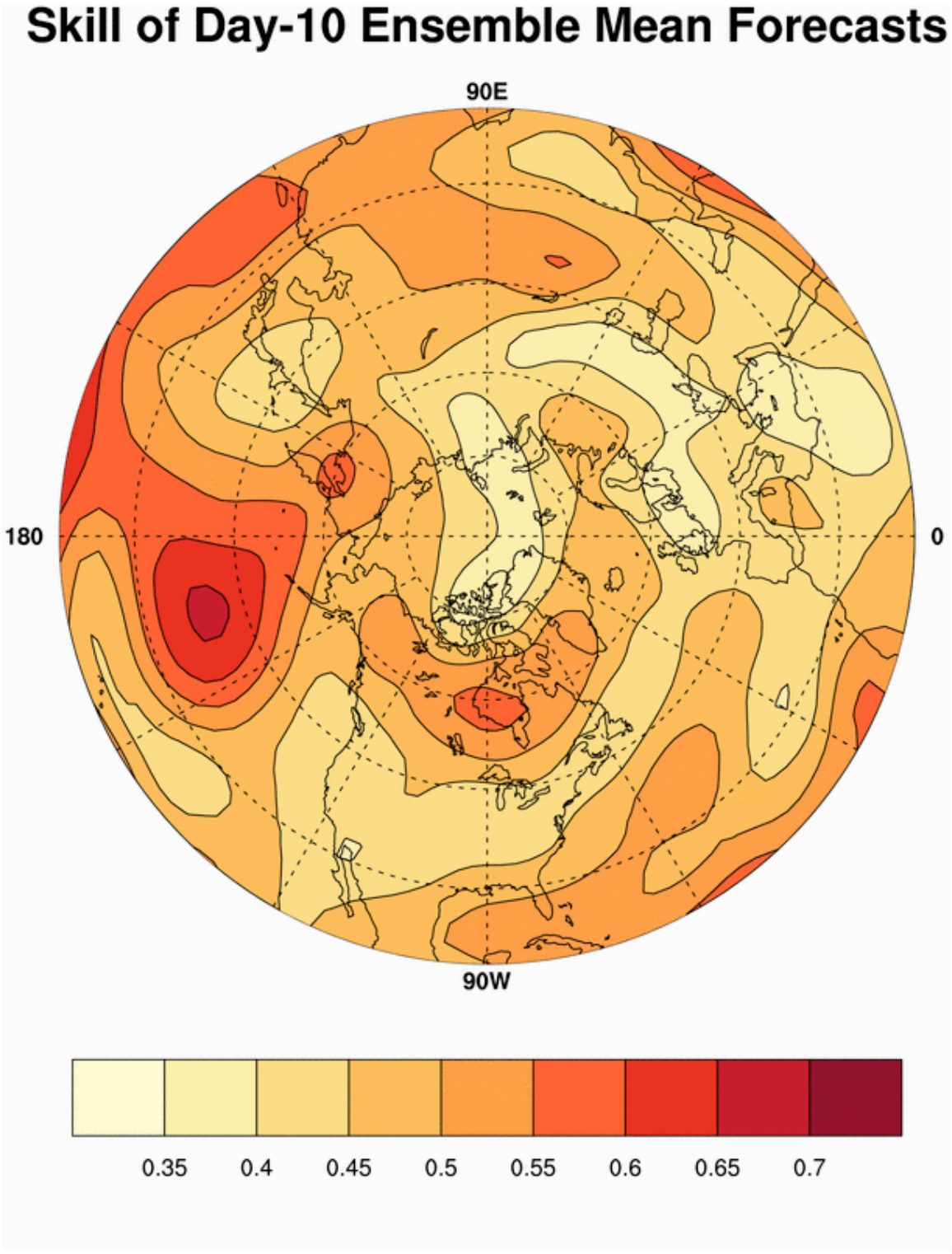

Figure 10: Correlation between time series of ensemble mean day-10 forecasts and corresponding verifying analyses (from the NCEP/NCAR reanalysis) at every grid point in the Northern Hemisphere for December to February 1979-2003.

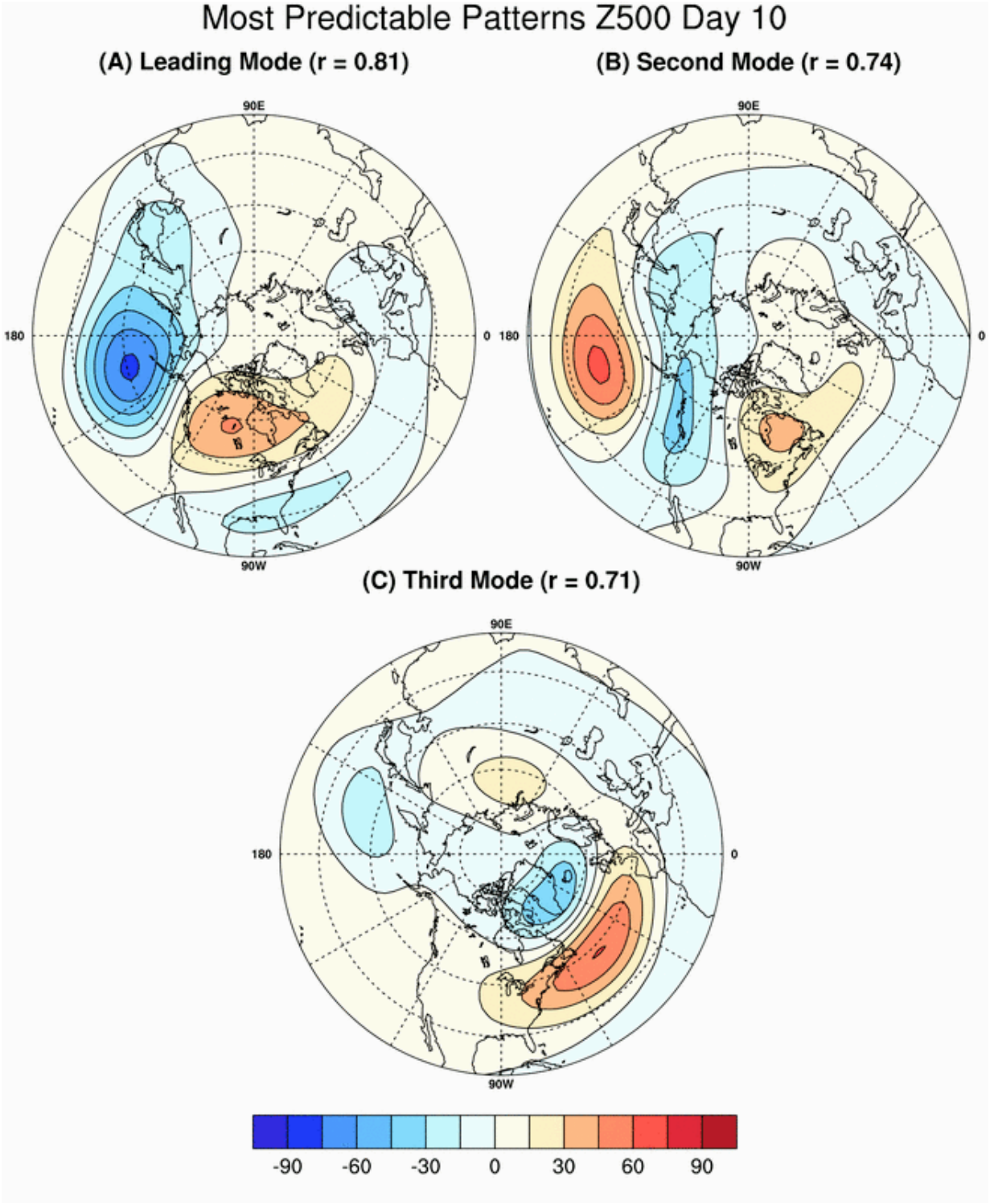

Figure 11 : Ensemble mean day-10 forecast 500 hPa height regressed on to the timeseries of the three most predictable forecast patterns. Contour interval 15 m. The correlation of between the time series of the predictor pattern and the corresponding predictand pattern (*r*) is given for each pattern.

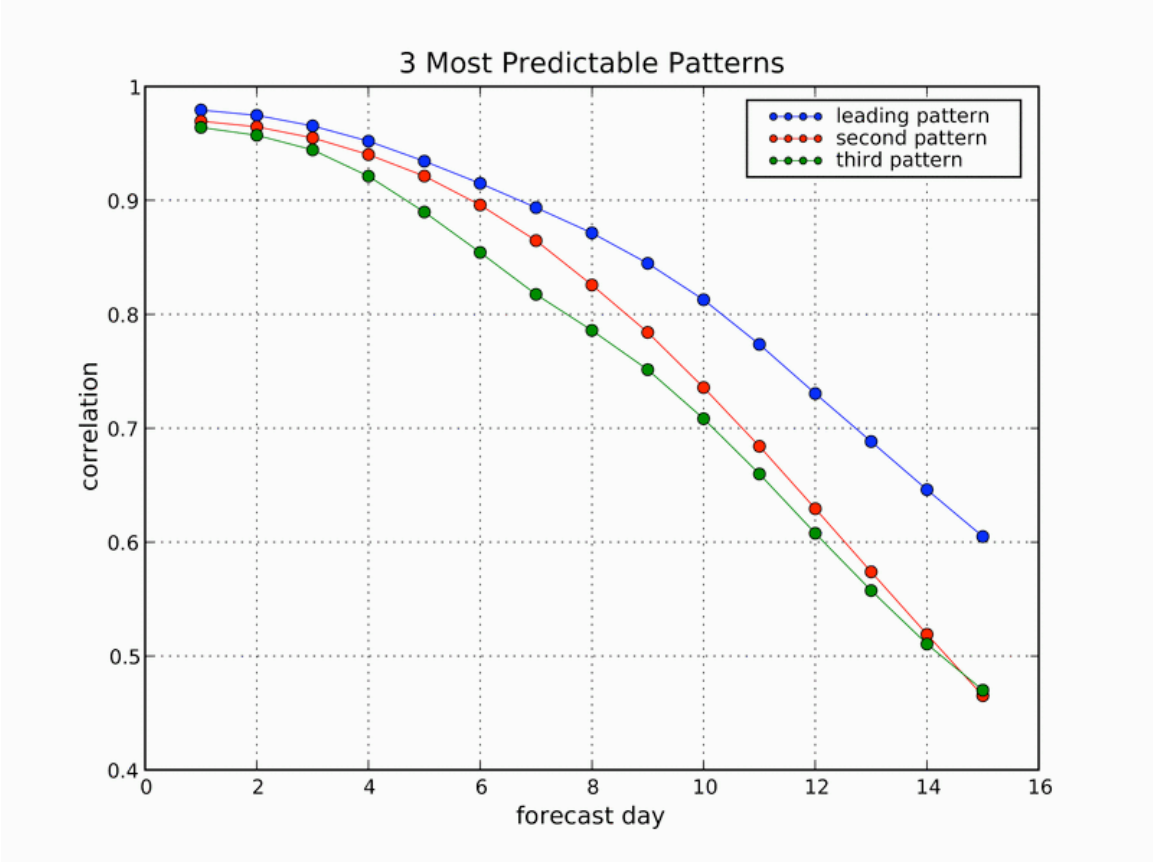

Figure 12: Correlation between the time series of the first three most predictable 500 hPa height day 10 forecast patterns and the time series of the corresponding patterns in the verifying analyses as a function of forecast lead time.